Socialist Federal Republic of Yugoslavia

FEDERAL PATENT OFFICE

[Coat of Arms]

The RO* "Center for Genetic Engineering", Scientific Institute for Molecular Biology, Vojvode Stepe [street] 233, 11000 Belgrade, has filed with the Federal Patent Office an addition to a patent application requesting patent protection in the Socialist Federal Republic of Yugoslavia for an invention related to:

"PROCESS FOR SEQUENCING GENOMES BY HYBRIDIZATION WITH OLIGONUCLEOTIDE PROBES"

This addition to the application with enclosures thereto was received on September 18, 1987 and was registered under No. 18617/87-P-570/87.

The Federal Patent Office hereby certifies that the enclosed copy of the said application and the enclosures thereto conform to the original.

Granted by the Federal Patent Office, after receiving payment of the correct fee, on February 16, 1988 under No. 1932/88 in Belgrade.

[Wax seal]

[Rubber stamp:]
Federal Patent Office,
Belgrade

By authorization of
the Director
[Signed:]
Rada Vukovic

*) Translator's Note: RO = Radnicka Organizacija = Workers' Organization

RO CENTER FOR GENETIC ENGINEERING

Addition to Patent Application No. 6425 of April 1, 1987.

This addition represents a further optimization of the sequences, of the number of required oligonucleotide probes and of the number and the length of cloned fragments of genomic DNA in order to achieve as complete a sequence as possible. Said addition does not change the essence of the method of achieving this or the determination of sequences of replicated fragments of genomic DNA and thus also of the sequences of the entire genomic DNA or parts thereof of one or several species by arranging positively hybridized oligonucleotide probes via overlapping sequences and by carrying out the hybridization under conditions in which these probes hybridize only to fully homologous sequences.

Because this addition includes several changes and clarifications, for greater clarity the addition is written in the form of a completely new DESCRIPTION which consists of 8 [Serbo-Croatian] pages. In the new description, we did not include the use of plasmid vectors and of amplification as possible methods for obtaining replicated fragments of genomic DNA. Competitive hybridization is the only substantially new part of the entire procedure. Moreover, the PATENT CLAIMS have been modified and now consist of 5 claims, and the ABSTRACT is new.

Applicant:
[Signed:]
Prof. Dr. Vladimir Glisin
Director

DESCRIPTION

Compared to prior-art sequencing methods, our procedure is based on an entirely different logic and is applicable only to the determination of sequences of complex DNA fragments and/or molecules (more than one million base pairs). It is based on highly specific hybridization of oligonucleotide probes (ONPs) with a length of 11 to 20 nucleotides.

Conditions for ONP hybridization have been found which differentiate between complete homology with a target and nonhomology in a single base [Wallace, R.B., et al., Nucleis Acids Res. 6, 3543-3557 (1979)]. When the hybridization method with 3 M tetramethylammonium chloride is used, the melting point of the hybrid depends only on the ONP length, regardless of the GC composition [Wood, W.I. et al. Proc. Natl. Acad. Sci. USA 82, 1585-1588 (1985)]. Hence, by hybridization under these conditions, sequences are determined unambiguously. By hybridization of genomic DNA, replicated in subclones (SC) of appropriate length, with a sufficient number of ONPs and by computerized arrangement of the detected sequences, it is possible to sequence the entire genome at the same time. We believe that this procedure is by an order of magnitude faster and less expensive than the one now being developed and that for this reason it is applicable to the sequencing of genomes of all characteristic species.

For this procedure, it is necessary to optimize the sequence length, the number of ONPs, the number of SCs and the length of the pooled DNA that can represent a hybridization spot.

11-meric ONPs are the shortest ONPs that can currently be successfully hybridized. This would *a priori* mean that 4^{11}, or 4194304, ONPs are necessary to detect each sequence. The same number of independent hybridizations would be required for each SC or SC pool (group). Positively hybridizing ONPs would arrange themselves over overlapping 10-mers. In this manner, the DNA sequence of the given SC would be obtained.

The process of SC sequence arrangement is interrupted when the overlapping 10-mer is repeated in the given SC. In this manner, uninterrupted sequences are obtained only between repeated 10-mers or longer oligonucleotide sequences (ONSs). These SC sequence fragments (SFs) cannot always be arranged into an unambiguous linear array without additional information. For this reason, it is important to determine the probable number of SFs (Nsf) for a given DNA length by use of probability calculations.

ONSs are distributed in a randomly formed long DNA sequence according to a binomial distribution. The average distance (A) between identical neighboring ONSs depends only on the ONS length (L) and is obtained from the expression $A = 4^L$. The probability that the ONS will be repeated N times in a fragment having a length of Lf base pairs (bp) is given by the equation

$$P(N,Lf) = C(N,Lf) \times (1/A)^N \times (1 - 1/A)^{Lf} \qquad (1)$$

wherein C(N,Lf) is the number of combinations of class N of Lf elements. The expected number of different ONSs of length L or the average distances A that repeat themselves N times in the fragment Lf bp is given by the product P(N,Lf) x A. If the arrangement of sequences is done via an overlapping ONS of length L, or average distance A_o, then Nsf in the fragment Lf is given by

$$Nsf = 1 + A_o \times \Sigma N \times P(N,Lf), \quad N \geq 2 \qquad (2)$$

In the event that all (4×10^6) 11-mers are used, about 3 SFs are expected in a 1.5 kb-long Lf. We shall return to the problem of SF arrangement later.

The number of 4×10^6 syntheses to obtain all 11-meric ONPs is uneconomical for the practical use of sequencing by hybridization (SBH). Deleting a significant number of ONPs (more than 20%) is not advantageous because it leads to unread gaps in the sequence. A much better method for reducing the number of independent ONP syntheses and independent hybridizations is by use of arranged ONP groups. In this case, shorter fragments must be sequenced, but there are no gaps in the sequence. The number of syntheses and hybridizations is reduced 40-fold, but 7 times more SCs are needed.

From the standpoint of information, the use of arranged ONP groups is the same as the use of shorter ONPs. For example, there are 65,536 different 8-meric ONSs. Since according to our current knowledge an ONS 8-mer cannot form a stable hybrid, a group of 11-mers can be used as an equivalent. Common to all 11-mers in the group is one 8-mer, so that information is obtained only about its presence or absence in the target DNA. The anticipated groups of 11-mers each contain 64 ONPs of the type (N2)N8(N1) (the 5',3' orientation is in the writing direction. (Nx) denotes x unspecified bases and Ny denotes y specific bases). With about 65,000 such groups, all sequences are detected. Based on Eq. 2, we find that an average of 3 SFs is expected in 200 bp-long DNA fragments. Because of variability, some fragments of this length will have 10 or more SFs.

ONPs of type (N2)N6(N1) are not very suitable for sequencing mammalian DNA because of the nonrandom GC and dinucleotide composition of this DNA. The common sequence of the ONP group must be l nger if it

contains more AT bases. Taking this into consideration, it is advantageous to use three kinds of probes: (N1)N10 where N10 are all the 10-mers not containing G and C, (N1)N9(N1) where N9 are all the 9-mers with 1 or 2 C÷G and (N2)N8(N1) where N8 are all the 8-mers containing 3 or more C÷G. About 81,000 such ONP groups are needed. Their average A_1 (Aa_1) is about 30,000. For the same A_1 value in random DNA, about 130,000 ONPs of type (N2)N8(A or T) and (N2)N8(or rG) are required. The given ONP groups permit the sequencing of 300-bp fragments with an average of 3 SFs. As a result of this 25% increase in the number of syntheses, the number of required SCs is reduced several times (to be discussed later).

In addition to these probes for solving problems of monotonic sequences, for confirmation of insert terminals and for supplementing information that is lost because of the impossibility of using ONPs that hybridize with the vector DNA, it is necessary to synthesize an additional 20,000 ONPs.

Monotonic sequences or, in general, one or two bp-long ONSs repeated in tandem (AAAAAAA.... TCTCTCTCTCT.... TGATGATGATGATGA...) represent a problem in SBH. With the above probes, it is not possible to determine the length of monotonic sequences that are longer than the common part of the ONP group. For this reason, for accurate determination of the length of monotonic ONSs that are up to 18 bp long, the following ONPs must be used: 16 An and Tn ONPs where n denotes 11 to 18 bp, 20 Cn and Gn ONPs where n denotes 9 to 18 bp, 4 (AT)n ONPs where n has the values (12, 14, 16, 18), 25 (AC)n, (AG)n, (TC)n, (TG)n and (CG)n ONPs where n has the values (10, 12, 14, 16, 18), 60 ONPs of type (N1N2N3)n which include all trinucleotides and n has the values (12, 15, 18), 180 ONPs of the type (N1N2N3N4)n which include all 4-mers and n has the values (12, 16, 18), 408 ONPs which include all 15 bp and 18 bp-long tandem 5-mers, 672 ONPs consisting of 18 bp long tandem 6-mers and 2340 ONPs consisting of 18 bp-long tandem 7-mers. The total number of these ONPs is 3725.

For the confirmation of the ends of DNA inserts in an SC, it is necessary to synthesize an additional 2048 ONPs of the N6(N5) or (N5)N6 type, where N6 denotes terminal vector sequences and (N5) denotes all the 5-mers in both cases.

The problem of vector DNA can be solved in two ways. One consists of prehybridization with cold vector DNA which is 7 bp shorter on both sides of the cloning site. The other method consists of leaving out ONPs that are complementary to the vector DNA. Because phage vector M13 was chosen as the most advantageous one (to be discussed later), about 7000 proposed ONPs will not be used. This is a significant percentage (11% of about 65,000 (N2)N8(N1) ONPs). This number can be reduced to about 3% by using, in place of the given 7000 ONPs,

an additional 21.000 ONPs of the (N1)N°1)N8(N1) type, where N8 denotes 7000 M13 8-mers and (N°1) denotes each of 3 nucleotides that is not present next to the given 8-mer.

Our calculations to date refer to sequencing single-stranded DNA. For the sequencing of double-stranded DNA, it is not necessary to synthesize both complementary ONPs. The number of required ONPs is thus halved. Because of the convenience of the M13 system, however, we will stay with the sequencing of single-stranded DNA. In this case, the use of one half of the ONPs in the SCs will lead to gaps of unread sequences. A gap in a given SC, however, will be read in the SC containing the complementary strand. In a representative random SC library, each sequence is represented on average 10 times. Hence, the probability exists that each sequence will be cloned in both senses, namely that both DNA strands will be read. It is thus possible to use only noncomplementary ONPs with an increased use of computerization. This means that the total number of required ONPs would be about 50.000. If it were possible to construct an M13 vector that could simultaneously or successively pack both strands, the use of noncomplementary ONPs would result in no additional requirements.

All SCs and/or SC groupings (SC pools) hybridize with all anticipated ONPs. In this manner, for each SC or SC pool we obtain a set of positively hybridizing ONPs. These ONPs are arranged in sequences by overlapping over the common sequences, which are only one nucleotide shorter than the ONP. For faster detection of overlapped ONPs in each synthesized ONP, it is necessary to determine in advance which ONPs show maximum overlap with it. Thus, each ONPx will have its subset of ONPs (ONPa, ONPb, ONPc, ONPd) 5' ONPx 3' (ONPe, ONPf, ONPg, ONPh). The arrangement is thus achieved by detecting which of the four ONPs with 5' and which of the four ONPs with 3' hybridized positively to the given SC or SC pool. The arranging continues until two positive overlapped ONPs are found for the last ONP arranged. When all SFs are extended to a maximum, this computer-assisted process ends.

By use of the given ONP groups, the number of SFs is increased by the given DNA length. In the general case, unambiguous arrangement is possible for a maximum of 3 SFs per SC, counted by the method by which Nsf was calculated by Eq. 2. Two of these are recognized as the terminal ones and the third is logically in the middle. The arrangement of SFs cannot be resolved by a suitable SC length, because it would be too short. Our solutions are: mutual arrangement of SFs and a large number of SCs so that the SC pools, too, can be used as a hybridization spot, and/or competitive hybridization of labeled and unlabeled ONPs.

To obtain as complete sequences as possible by SBH, SCs that are to be used later require three SC libraries in the M13 vector with inserts of 0.5 kb and 7 kb as well as different inserts of different sizes made up of two sequences which in the genomic DNA are separated by about 100 kb (skipping the SC). The first library serves primarily for arranging the SFs. These SCs can also be kept for later experimental use. These SCs

participate in the hybridization as pools obtained during phage growth by simultaneous infection or after the growth. The second library is the basic one. Its SCs with their larger inserts are more suitably stored for further use. The 7 kb length was chosen as the upper limit for the size of inserts that can successfully be cloned in M13. The third library serves to correctly link into a single sequence parts of sequences separated by highly homologous sequences longer than 7 kb and by uncloned DNA fragments.

Following hybridization of SCs of all libraries with all ONPs and after SF computing, the mutual arrangement of SFs and SCs is undertaken. The basic library is arranged first. The overlapped SCs are detected through the content of the entire starting SF of the starting SC or parts thereof. Suitable for mammals are, in the general case, all SFs with a length of about 20 bp and longer. The average SF length of these SCs was calculated by Eq. 2 and found to be from 2 to 12 bp. This indicates the existence of a sufficient number of SFs of suitable length. Moreover, these SFs, of which mostly there are two, are known, and one of them follows the starting SF. In this case, both sequences are examined, one of them being the right one, and the overlapped SCs are detected via this sequence. The exact displacement of the overlapping SCs relative to the starting SC is determined on the basis of the remaining SF content. At the same time, by detecting all SCs that overlap with the starting SC, the SFs of the starting SC are grouped into a linear array of subsets (SSF). The SSFs are defined by neighboring endings of overlapped SCs (start-start, start-end or end-end). The SC overlapping process continues via the SF taken from the most protruding SSF of the most protruding SC. The arranging process is interrupted when the uncloned part of the DNA is encountered or, as in SF formation, when a repeat sequence longer than 7 kb is encountered. This procedure affords maximum-size groups of arranged, overlapped 7-kb-long SCs and linearly ordered SSFs of their SFs.

In this procedure for arranging SFs, the DNA length that includes the SSFs is essential. This length depends on the number of SSFs, which is equal to the number of SC endings, namely it is twice as large as the number of SCs. For a representative library of DNA fragments of one million bp, 700 7-kb-long SCs are needed. This means that the average SSF size is 715 bp. The actual average number of SFs within such an SSF is not even one tenth of all SFs of the entire 7-kb SC. The actual number is independent of the SC length, namely it depends only on the SSF length. According to Eq. 2, for a length of 715 bp and an Aa_0 of 30,000 that the anticipated ONPs have, the expected average number of SFs with an average length of 45 bp is 16.

The arrangement of SFs within the SSFs obtained is accomplished via a 0.5-kb SC library. In this procedure, it is not essential that these be individual SCs; an SC pool can also be used. The SCs in a pool are informative if they do not overlap with each other. From an information and technology standpoint, a 10-kb pool of cloned DNA is advantageous, although it does not represent a limit. The required number of these SCs or pools is such that the maximum size of the SSF they form will not be greater than 300 bp. With the proposed ONP, we

anticipate within this DNA length 3 SFs (Eq. 2), which, as already explained, can be arranged unambiguously. By use of the binomial distribution, we derived the equation

$$Nsa = 2Nsc\,(1 - 2Nsc/Nbp)^{Lms} \qquad (3)$$

wherein Nsa is the number of SSFs greater than Lms, Nsc is the number of SCs, Nbp is the number of base pairs in the DNA fragment or molecule being sequenced, Lms is the SSF size that, on average, gives the arrangement number of 3 SF and which in this case is 300 bp. On the basis of this equation, it was determined that 25,000 0.5-kb SCs are needed for a DNA fragment of one million bp. The number of 10-kb pools is 1250. The average SSF size obtained for this SC is 20 bp.

SF arranging is done by computer-assisted detection of pools containing SCs that overlap the starting SSF obtained by arranging the basic library. The detection is performed on the basis of the content of the entire randomly selected SF or part thereof in the starting SSF. Based on the content of the other SFs of the starting SSF, the size of the overlap of 0.5-kb SCs is determined, and at the same time, because of their high density, the order of SFs in the starting SSF is also determined. At the end of this process, one obtains the sequence of each group of arranged 7-kb SCs and an indication as to the pool that contains the 0.5-kb SC that contains the particular sequence. At a certain small number of locations, the sequence will not be complete or it will be ambiguous. Our calculations show that this happens on average at less than one location per million bp, the randomly distributed undetected ONSs amounting to 30%. These locations are sequenced by suitable treatment of the SCs containing them and repeated application of SBH or by competitive hybridization of suitably selected pairs of unlabeled and labeled ONPs or by the conventional method or by the advanced conventional method.

The competitive hybridization procedure will be explained on the example of a twice repeated 7-bp sequence. In this case, two SFs terminate and two start with the repeating sequence TTAAAAGG, which is underlined:

5'NNNNNNNNNNNNCATTAAAAGG3'
5'NNNNNNNNNNNNCGTTAAAAGG3'

5'TTAAAAGGTACNNNNNNN3'
5'TTAAAAGGCCGNNNNNNN3'

By prehybridization with excess unlabeled ONP, for example 5'(N2)CATTAAAAG(N1)3' which because of a noncomplementary base cannot hybridize to 5'NNCGTTAAAAGG3', the subsequent hybridization of one of the two labeled ONPs, i.e., 5'(N2)AAAAGGTAC(N1)3' or 5'(N2)AAAAGGCCG(N1)3', is prevented. The pair

of probes that compete with each other defines the SFs that continue one after the other. This can be confirmed by an alternative selection of a suitable pair of ONPs. This procedure can be applied to all repeating ONSs with length up to 18 bp. To apply the procedure to the arranging of the large number of SFs, the prehybridization must be separated from the hybridization in space and time. For this reason, the stability of the unlabeled ONP is important. If stability cannot be achieved via ONP concentrations and hybridization temperatures, the cold ONPs will covalently bind to the complementary DNA that is bound to the filter by exposure to UV radiation in the presence of psoralen or by use of ONPs bearing a reactive group capable of covalent binding.

The SCs of the third library are used to bind the sequenced parts into a single DNA fragment. About 170 SCs are needed for one million bp. For larger DNA fragments, the values are directly proportional to this and other numbers calculated for one million bp. Because these SCs contain sequences that are separated by an average of 100 kb, with them it is possible to skip repeated or uncloned sequences that are up to 100 kb long by finding out which two sequenced parts contain sequences that are present in an SC of this library.

The experimental requirements of this method are represented by the total number of ONPs, the hybridization of 50,000 thereof and to the number of separate SCs out of 2120 that must be hybridized for a 1-million bp DNA fragment.

The described libraries are created in phage vector M13. This vector facilitates the cloning of 100-7000 bp-long inserts and gives a high titer of recombinant phages from bacterial cells by budding without cell lysis. Centrifuging the bacterial cultures gives a phage preparation that is not contaminated with bacterial DNA, and the bacterial sediment can be used for another phage production after adding the nutrient medium. By addition of alkali, DNA separates from the protein envelope and at the same time undergoes denaturation. This results in effective spotting and covalent binding to nylon filters on which the hybridization is carried out. Hybridization of one SC with all ONPs requires an amount of DNA equal to that which can be obtained from a few milliliters of bacterial culture. Most advantageous for growth and robotic application to filters are plates that are similar to microtitration plates and have appropriate dimensions and holes of appropriate volume.

The DNA is applied to the filters with a robotic arm. An arm with 10,000 suction apertures is sufficient for sequencing even the larges genomes. After aspiration of the DNA solution from the holes of the "micro" plates, the suction elements are brought closer to each other by means of a reducing head until they are separated by a distance of 1 mm. Then, an appropriate amount of DNA and at the same time 10,000 SCs are applied to the filter. This is repeated on the required number of filters with the same 10,000 SCs. The same is then done with all the other SCs in groups of 10,000. The number of "prints" of a group of 10,000 SCs for 50,000 ONPs is about 1000, because each filter can be washed and reused 50 times.

Hybridization is carried out in cycles. One cycle requires one day at the most. During one cycle, all SCs are hybridized with the defined number of ONPs. To complete all hybridizations within a reasonable length of time, about 1000 containers with one ONP each are used per cycle. To save on ONPs, a smaller volume of hybridization liquid is used, and the filters are added in several steps. The filters from all hybridization containers are collected in one container where all of them are subjected to further treatment at the same time, namely to washing and to the performance of color reactions when biotinylized rather than radioactively labeled ONPs are used. By carrying out the hybridization in 20 x 20 x 20 cm containers and without repeating the individual cycle, the SCs required for sequencing DNAs as large as 10^{10} bp can be hybridized.

After each hybridization cycle, the technological procedure continues by reading the hybridization results. The data are stored in the computer memory of the computer center. The data are of a binary character (+,-) and their reading involves several sensitivity thresholds. From these data, at the computer center, the SFs are arranged first followed by the mutual arrangement of the SFs and SCs. After all data have been processed, the computer center determines which SC must be subjected to what additional treatment to obtain the complete sequence.

SBH is a method that minimizes experimental work at the expense of more computer work. The only techological requirement is sequencing by specific ONP hybridization. The non-use of up to 6% of the envisaged ONPs can be tolerated without the appearance of gaps in the reading of the DNA sequence. To reduce the number of false negatives (unsuccessful hybridization of ONPs because of their limited hybridization length of 11 nucleotides) and to eliminate false positives, the envisaged ONPs have nonspecified bases at the ends, namely at the only locations where errors are possible. In place of ONPs of group (N3)N8, ONPs of group (N2)N8(N1) are used. For this reason, even the ONPs meant for measuring the length of monotonic sequences are synthesized as (N1)Nx(N1) ONP groups. In the case of certain basic ONP groups which give many false negatives, ONP groups of the (N2)N8(N2) type are used, and the hybridization is carried out at the temperature used for 11-meric ONPs.

The formation of internal duplexes in the DNA that is bound to the filter is one of the known structural reasons for false negatives or gaps in sequence reading. This problem is overcome by improving the bonding of DNA to nylon filters and by cutting the DNA into fragments of an average size of 50 bp (ultrasound, acid, endonuclease) before applying it to the filter. A significant number of these fragments of the recombinant molecule will also be cut within the duplex structure. The formation of such a structure is thus prevented making hybridization possible.

The approach on which this procedure is based makes it possible for a computer-controlled, fully robotized line to produce sufficient data in the form of binary signals from which the sequence of complex DNA fragments or molecules can then be obtained by computing.

PATENT CLAIMS

1. Procedure for sequencing the entire genomic DNA or large parts thereof by hybridization with oligonucleotide probes, characterized in that the replicated fragments of genomic DNA are hybridized with some or all 8- to 20-nucleotides-long ONPs resulting from the variation and repetition of the 4 nucleotides A, T or U, C, G, or their derivatives and analogs, by using individual ONPs or a mixture of individually synthesized ONPs, or of an arranged ONP group synthesized so that more or all nucleotides or their derivatives or analogs are added at a certain point during synthesis, and that the hybridization reaction is carried out under conditions in which the oligonucleotide probes hybridize only with a fully homologous sequence or a sequence that has an amount of nonhomology that does not cause the formation of ambiguous or faulty sequences in the process of arranging positively hybridizing ONPs via a maximum mutual overlap of their sequences.

2. Procedure according to Claim 1, characterized in that replicated fragments of genomic DNA are obtained by cloning into vectors based on single-stranded bacteriophages in the form of three subclone libraries with inserts of 0.1 to 1 kb and 3 to 10 kb and inserts consisting of two parts separated from the genomic DNA by an average of 50 to 200 kb, that they are replicated as individual subclones and as SC groups obtained by simultaneous infection and that they are hybridized on the filter to which they are applied as a hybridization spot as uninterrupted or cut out vector-insert DNAs of individual subclones and groups of subclones up to fragments of an average length of 20 bp.

3. Procedure according to Claims 1 and 2, characterized in that the subfragments of the sequence of the individual subclones or groups of subclones, obtained by overlapping positively hybridizing ONPs for the given subclone or group of subclones, are arranged into a natural linear array by cyclic detection of overlapping subclones based on the content of subfragments of the sequence of the starting subclone or group of subclones, which subclones in a library of 0.1 to 1 kb show an average displacement of less than 100 kb.

4. Procedure according to Claims 1 and 2, characterized in that the subfragments of the sequence of the individual subclone or group of subclones, obtained by overlapping positively hybridizing ONPs for the given subclone or group of subclones, are arranged into a natural linear array by the procedure of competitive hybridization with unlabeled and labeled oligonucleotide probes whereby first the filter hybridizes with a saturating amount of unlabeled oligonucleotide probe, which contains all or part of the terminal, repeating oligonucleotide sequences in the sequence subfragment for which it is desired to determine the following sequence subfragment and then, with or without previous covalent bonding of this cold probe to the filter, separate hybridizations are carried out with labeled oligonucleotides with probes containing all or part f the repeating ligonucleotide sequence, so

that at least a part is common to that part of the repeating sequence which is contained in the unlabeled probe, and the remainder of the nonrepeating sequences that follow the repeating sequence, from each sequence subfragment that contains the repeating sequence, and the following sequence subfragment is determined as that whose labeled oligonucleotide probe does not hybridize.

5. Procedure according to Claims 1-4, characterized in that the sequencing of a million bp mammalian DNA is carried out with individual hybridization spots containing 1250 groups of 20, or an average of 20, groups of M13 subclones 0.5 kb in length, 700 M13 subclones 7 kb in length and 170 skipping M13 subclones which skip on average 100 kb of the genomic DNA, and by hybridization of each spot with 1024 groups of 16 probes each of the (A,T,C,G)N10(A,T,C,G) type wherein N10 are all the 10-mers that do not contain the G and C nucleotides, with 23,040 groups of 16 probes each of the (A,T,C,G)N9(A,T,C,G) type where N9 are all the 9-mers containing one or two C+G nucleotides, with 55,834 groups of 64 probes each of the (A,T,C,G)(A,T,C,G)N8(A,T,C,G) type or of the (A,T,C,G)(A,T,C,G)N8(A,T,C,G)(A,T,C,G) type where N8 are all the 8-mers containing three or more C+G nucleotides, and with 3725 groups of 16 probes each of the (A,T,C,G)Nm(A,T,C,G) type where Nm denotes all monotonic sequences of the required lengths shorter than 18 bp and consisting of 1 to 7 nucleotides-long repeating units.

Applicant
[Signed:]
Prof. Dr. Vladimir Glisin,
Director

ABSTRACT

The conditions under which oligonucleotide probes hybridize only with fully homologous sequences are known. By such hybridization and by arranging positively hybridizing probes via overlapping parts, the sequence of the given DNA fragment is read. By simultaneous hybridization of DNA molecules of the single-stranded phage vector-cloned insert, applied in the form of spots, with about 50,000 to 100,000 groups of probes the main type of which is (A,T,C,G)(A,T,C,G)N8(A,T,C,G), information for computer-assisted determination of DNA sequences of the complexity of the mammalian genome can be obtained. To obtain as complete sequences as possible, three libraries in the vector based on the M13 phage are used: those with 0.5 kb inserts, those with 7 kb inserts and those with inserts consisting of two sequences separated in the genomic DNA by an average of 100 kb. For one million bp of genomic DNA are needed 25,000 0.5-kb subclones, 700 7-kb subclones and 170 skipping subclones. The 0.5 kb subclones are applied to the filter in groups of 20, so that the total number of samples is 2120 per million bp. The procedure can be readily and completely robotized for reading complex genomic DNA fragments or molecules in a manufacturing plant.